May 1990

LIDS-P-1977

Metropolis-type Annealing Algorithms for Global Optimization in $\mathbb{R}^d$

by

Saul B. Gelfand[1] and Sanjoy K. Mitter[2]

Abstract

We establish the convergence of a class of Metropolis-type Markov chain annealing algorithms for global optimization of a smooth function $U(\cdot)$ on $\mathbb{R}^d$. No prior information is assumed as to what bounded region contains a global minimum. Our analysis is based on writing the Metropolis-type algorithm in the form of a recursive stochastic algorithm $X_{k+1} = X_k - a_k(\nabla U(X_k) + \xi_k) + b_k W_k$, where $\{W_k\}$ are independent standard Gaussian random variables, $\{\xi_k\}$ are (unbounded, correlated) random variables, and $a_k = A/k$, $b_k = \sqrt{B}/\sqrt{k \log\log k}$ for k large, and then applying results about $\{X_k\}$ from [15]. Since the analysis of $\{X_k\}$ in [15] is based on the asymptotic behavior of the related Langevin-type Markov diffusion annealing algorithm $dY(t) = -\nabla U(Y(t))dt + c(t)dW(t)$, where $W(\cdot)$ is a standard Wiener process and $c(t) = \sqrt{C}/\sqrt{\log t}$ for t large, this work demonstrates and exploits the close relationship between the Markov chain and diffusion versions of simulated annealing.

Key words: global optimization, random optimization, simulated annealing, stochastic gradient algorithms, Markov chains.

---

[1] Computer Vision and Image Processing Laboratory, School of Electrical Engineering, Purdue University, West Lafayette, IN 47907.

[2] Laboratory for Information and Decision Systems and Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139. Research partially supported by the Air Force Office of Scientific Research grant AFOSR 89-0276.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
MAY 1990		**00-05-1990 to 00-05-1990**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Metropolis-type Annealing Algorithms for Global Optimization in IRd	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Massachusetts Institute of Technology,Laboratory for Information and Decision Systems,77 Massachusetts Avenue,Cambridge,MA,02139-4307**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**30**	

1. INTRODUCTION

Let $U(\cdot)$ be a real-valued function on some set Σ. The global optimization problem is to find an element of the set $S^* = \{x: U(x) \leq U(y)\ \forall\ y \in \Sigma\}$ (assuming $S^* \neq \phi$). Recently, there has been alot of interest in the simulated annealing method for global optimization. Annealing algorithms were initially proposed for finite optimization (Σ finite), and later developed for continuous optimization ($\Sigma = \mathbb{R}^d$). An annealing algorithm for finite optimization was first suggested in [1], [2], and is based on simulating a finite-state Metropolis-type Markov chain. The Metropolis algorithm and other related algorithms such as the "heat bath" algorithm, were originally developed as Markov chain sampling methods for sampling from a Gibbs distribution [3]. The asymptotic behavior of finite state Metropolis-type annealing algorithms has been extensively analyzed [4]-[9].

A continuous time annealing algorithm for continuous optimization was first suggested in [10], [11] and is based on simulating a Langevin-type Markov diffusion:

$$dY(t) = -\nabla U(Y(t))dt + c(t)dW(t)\ . \tag{1.1}$$

Here $U(\cdot)$ is a smooth function on $\mathbb{R}^d$, $W(\cdot)$ is a standard d-dimensional Wiener process, and $c(\cdot)$ is a positive function with $c(t) \longrightarrow 0$ as $t \longrightarrow \infty$. In the terminology of simulated annealing algorithms, $U(x)$ is called the energy of state x, and $T(t) = c^2(t)/2$ is called the temperature at time t. Note that for a fixed temperature $T(t) = T$, the resulting Langevin diffusion like the Metropolis chain has a Gibbs distribution $\propto \exp(-U(x)/T)$ as its unique invariant distribution. Now (1.1) arises by adding slowly decreasing white Gaussian noise to the continuous time gradient algorithm

$$\dot{z}(t) = -\nabla U(z(t))\ . \tag{1.2}$$

The idea behind using (1.1) instead of (1.2) for minimizing $U(\cdot)$ is to avoid getting trapped in strictly local minima. The asymptotic behavior of $Y(t)$ as $t \longrightarrow \infty$ has been studied in [10], [12]-[14]. In [10], [14] convergence results were obtained for a version of (1.1) which was modified to constrain the trajectories to lie in a fixed bounded set (and hence is only applicable to global optimization over a compact subset of $\mathbb{R}^d$); in [12], [13] results were obtained for global optimization over all of $\mathbb{R}^d$. Chiang, Hwang and Sheu's main result from [12] can be roughly stated as follows: if $U(\cdot)$ is suitably behaved and $c^2(t) = C/\log t$ for t large with $C > C_0$ (a constant depending only on

$U(\cdot)$), then $Y(t) \rightarrow S^*$ as $t \rightarrow \infty$ in probability.

A discrete time annealing algorithm for continuous optimization was suggested in [14], [15] and is based on simulating a recursive stochastic algorithm

$$X_{k+1} = X_k - a_k(\nabla U(X_k) + \xi_k) + b_k W_k \ . \tag{1.3}$$

Here $U(\cdot)$ is again a smooth function on $\mathbb{R}^d$, $\{\xi_k\}$ is a sequence of $\mathbb{R}^d$-valued random variables, $\{W_k\}$ is a sequence of independent standard d-dimensional Gaussian random variables, and $\{a_k\}$, $\{b_k\}$ are sequences of positive numbers with $a_k, b_k \rightarrow 0$ as $k \rightarrow \infty$. The algorithm (1.3) could arise from a discretization or numerical integration of the diffusion (1.1) so as to be suitable for implementation on a digital computer; in this case ξ_k is due to the discretization error. Alternatively, the algorithm (1.3) could arise by artificially adding slowly decreasing white Gaussian noise (i.e., the $b_k W_k$ terms) to a stochastic gradient algorithm

$$Z_{k+1} = Z_k - a_k(\nabla U(Z_k) + \xi_k) \tag{1.4}$$

which arises in a variety of optimization problems including adaptive filtering, identification and control; in this case ξ_k is due to noisy or imprecise measurements of $\nabla U(\cdot)$ (c.f. [16]). The idea behind using (1.3) instead of (1.4) for minimizing $U(\cdot)$ is to avoid getting trapped in strictly local minima. In the sequel we will refer to (1.4) and (1.3) as standard and modified stochastic gradient algorithms, respectively. The asymptotic behavior of X_k as $k \rightarrow \infty$ has been studied in [14], [15]. In [14] convergence results were obtained for a version of (1.3) which was modified to constrain the trajectories to lie in a compact set (and hence is only applicable to global optimization over a compact subset of $\mathbb{R}^d$); in [15] results were obtained for global optimization over all of $\mathbb{R}^d$. Also, in [14] convergence is obtained essentially only for the case where $\xi_k = 0$; in [15] convergence is obtained for $\{\xi_k\}$ with unbounded variance. This latter fact has important implications when $\nabla U(\cdot)$ is not measured exactly. Our main result from [15] can be roughly stated as follows: if $U(\cdot)$ and $\{\xi_k\}$ are suitably behaved, $a_k = A/k$ and $b_k^2 = B/k \log\log k$ for k large with $B/A > C_0$ (the same C_0 as above), and $\{X_k\}$ is tight, then $X_k \rightarrow S^*$ as $k \rightarrow \infty$ in probability (conditions are also given in [15] for tightness of $\{X_k\}$). Our analysis in [15] of the asymptotic behavior of X_k as $k \rightarrow \infty$ is based on the asymptotic behavior of the associated SDE (1.1). This is analogous to the well-known method of analyzing the asymptotic behavior of Z_k as $k \rightarrow \infty$ based on the asymptotic behavior of the associated ODE (1.2) [16], [17].

It has also been suggested that continuous (global) optimization might be performed by simulating a continuous-state Metropolis-type Markov chain [10], [18], [19]. Although some numerical work has been performed with continuous-state Metropolis-type annealing algorithms there has been very little theoretical analysis, and furthermore the analysis of the continuous state case does not follow from the finite state case in a straightforward way (especially for an unbounded state space). The only analysis we are aware of is in [19] where a certain asymptotic stability property is established for a related algorithm and a particular cost function which arises in a problem of image restoration.

In this paper we demonstrate the convergence of a class of continuous state Metropolis-type Markov chain annealing algorithms for general cost functions. Our approach is to write such an algorithm in (essentially) the form of a modified stochastic gradient algorithm (1.3) for suitable choice of ξ_k, and to apply results from [15]. A convergence result is obtained for global optimization over all of $\mathbb{R}^d$. Some care is necessary to formulate a Metropolis-type Markov chain with appropriate scaling. It turns out that writing the Metropolis-type annealing algorithm in the form (1.3) is rather more complicated than writing standard variations of gradient algorithms which use some type of (possibly noisy) finite difference estimate of $\nabla U(\cdot)$ in the form (1.4) (c.f. [16]). Indeed, to the extent that the Metropolis-type annealing algorithm uses an estimate of $\nabla U(\cdot)$, it does so in a much more subtle manner than a finite difference approximation, as will be seen in the analysis.

Since our convergence results for the Metropolis-type Markov chain annealing algorithm are ultimately based on the asymptotic behavior of the Langevin-type Markov diffusion annealing algorithm, this paper demonstrates and exploits the close relationship between the Markov chain and diffusion versions of simulated annealing, which is particularly interesting in view of the fact that the development and analysis of these methods has proceeded more-or-less independently. We remark that similar convergence results for other continuous-state Markov chain sampling method based annealing algorithms (such as the "heat bath" method) can be obtained by a procedure similar to that used in this paper.

The paper is organized as follows. In Section 2 we discuss appropriately modified versions of the tightness and convergence results for modified stochastic gradient

algorithms as given in [15]. In Section 3 we present a class of continuous state Metropolis-type annealing algorithms and state some convergence theorems. In Section 4, we prove the convergence theorems of Section 3 using the results of Section 2.

2. MODIFIED STOCHASTIC GRADIENT ALGORITHMS

In this Section we give convergence and tightness results for modified stochastic gradient algorithms of essentially the type described in Section 1. The algorithms and theorems discussed below are a slight variation on the results of [15], and are appropriate for proving convergence and tightness for a class of continuous state Metropolis-type annealing algorithms (see Section 3,4).

We use the following notations throughout the paper. Let $\nabla U(\cdot)$, $\Delta U(\cdot)$, and $HU(\cdot)$ denote the gradient, Laplacian and Hessian matrix of $U(\cdot)$, respectively. Let $|\cdot|$, $<\cdot,\cdot>$ and $\otimes$ denote Euclidean norm, inner product, and outer product, respectively. For real numbers a and b let $a \vee b = \text{maximum}\{a,b\}$, $a \wedge b = \text{minimum}\{a,b\}$, $[a]_+ = a \vee 0$, and $[a]_- = a \wedge 0$. For a process $\{X_k\}$ and a function $f(\cdot)$, let $E_{n,x}\{f(X_k)\}$, $P_{n,x}\{f(X_k)\}$ denote conditional expectation and probability given $X_n = x$ (more precisely, these are suitable fixed versions of the conditional expectation and probability). Also for a measure $\mu(\cdot)$ and a function $f(\cdot)$ let $\mu(f) = \int f d\mu$. Finally, let $N(m,R)(\cdot)$ denote normal measure with mean m and covariance matrix R, and let I denote the identity matrix.

2.1. Convergence

In this subsection we consider the convergence of the discrete time algorithm[+]

$$X_{k+1} = X_k - a_k(\nabla U(X_k) + \xi_k) + b_k(|X_k| \vee 1)W_k \,. \tag{2.1}$$

Here $U(\cdot)$ is a smooth real-valued function on $\mathbb{R}^d$, $\{\xi_k\}$ is a sequence of $\mathbb{R}^d$-valued random variables, $\{W_k\}$ is a sequence of independent standard d-dimensional Gaussian random variables, and

[+] The results are not changed if we replace $|X_k| \vee 1$ by $|X_k| \vee a$ or $|X_k| + a$ for $a \geq 1$.

$$a_k = \frac{A}{k} \ , \ b_k = \frac{\sqrt{B}}{\sqrt{k \log\log k}} \ , \ k \text{ large} \, ,$$

where A, B are positive constants.

For $k = 0,1,\ldots$ let $\mathcal{F}_k = \sigma(X_0, W_0, \ldots, W_{k-1}, \xi_0, \ldots \xi_{k-1})$. In the sequel we will consider the following conditions (α, β are constants whose values will be specified later).

(A1) $U(\cdot)$ is a C^2 function from $\mathbb{R}^d$ to $[0,\infty)$ such that

$$\underline{\lim}_{|x|\to\infty} \frac{|\nabla U(x)|}{|x|} > 0$$

$$\lim_{|x|\to\infty} \left\langle \frac{\nabla U(x)}{|\nabla U(x)|} \ , \ \frac{x}{|x|} \right\rangle = 1$$

$$\inf_x \left(|\nabla U(x)|^2 - \Delta U(x) \right) > -\infty$$

(A2) For $\epsilon > 0$ let

$$d\pi^\epsilon(x) = \frac{1}{Z^\epsilon} \exp\left[-\frac{2U(x)}{\epsilon^2} \right] dx \ , \ Z^\epsilon = \int \exp\left[-\frac{2U(x)}{\epsilon^2} \right] dx < \infty \, .$$

π^ϵ has a weak limit π as $\epsilon \to 0$.

(A3) Let K be a compact subset of $\mathbb{R}^d$. Then there exists $L \geq 0$ such that

$$E\{ |\xi_k|^2 \, | \mathcal{F}_k \} \leq L a_k^\alpha \ , \ \forall \, X_k \in K \ , \ \text{w.p.1}$$

$$|E\{ \xi_k \, | \mathcal{F}_k \}| \leq L a_k^\beta \ , \ \forall \, X_k \in K \ , \ \text{w.p.1} \, .$$

W_k is independent of $\mathcal{F}_k$.

We note that π concentrates on S^*, the global minima of $U(\cdot)$. For example, if S^* consists of a finite number of points, then π exists and is uniformly distributed over S^*. The existence of π and a simple characterization in terms of $HU(\cdot)$ is discussed in [20].

In [12] and [15] it was shown that there exists a constant C_0 which plays a critical role in the convergence of (1.1) and (1.3), respectively (in [12] C_0 was denoted by c_0). C_0 has a interpretation in terms of the action functional for the dynamical system (1.2); see [12] for an explicit expression for C_0 and some examples. The constant C_0 plays the same role in the convergence of (2.1) considered here.

Let $K_1 \subset \mathbb{R}^d$ and let $\{X_k^x\}$ denote the solution of (2.1) with $X_0 = x$. We shall say that $\{X_k^x\colon k \geq 0,\ x \in K_1\}$ is tight if given $\epsilon > 0$ there exists a compact $K_2 \subset \mathbb{R}^d$ such that $P_{0,x}\{X_k \in K_2\} > 1 - \epsilon$ for all $k \geq 0$ and $x \in K_1$. Here is our theorem on the convergence of X_k as $k \to \infty$.

Theorem 1: Assume (A1), (A2), (A3) hold with $\alpha > -1$ and $\beta > 0$. Let $\{X_k\}$ be given by (2.1), and assume $\{X_k^x\colon k \geq 0,\ x \in K\}$ is tight for K a compact set. Then for $B/A > C_0$ and any bounded continuous function $f(\cdot)$ on $\mathbb{R}^d$

$$\lim_{k\to\infty} E_{0,x}\{f(X_k)\} = \pi(f)$$

uniformly for x in a compact set.

Note that since π concentrates on S^*, under the conditions of Theorem 1 we have $X_k \to S^*$ as $k \to \infty$ in probability.

Theorem 1 is proved similiarly to [15, Theorem 2] where we considered the algorithm

$$X_{k+1} = X_k - a_k(\nabla U(X_k) + \xi_k) + b_k W_k\ , \tag{2.2}$$

and we will not go through the details here. The main difference between the conditions and proofs of Theorem 1 and [15, Theorem 2] is that in Theorem 1 the condition $\underline{\lim}_{|x|\to\infty} |\nabla U(x)|/|x| > 0$ is needed to establish the tightness of $\{Y(t)\}$ for the diffusion $dY(t) = -\nabla U(Y(t))dt + c(t)(|Y(t)| \vee 1)dW(t)$ associated with (2.1), whereas in [15, Theorem 2] the weaker condition $\lim_{|x|\to\infty} |\nabla U(x)| = \infty$ suffices to establish the tightness of $\{Y(t)\}$ for the diffusion $dY(t) = -\nabla U(Y(t))dt + c(t)dW(t)$ associated with (2.2).

2.2. Tightness

In this subsection we consider the tightness of the discrete time algorithm[+]

$$X_{k+1} = X_k - a_k(\psi_k(X_k) + \eta_k) + b_k(|X_k| \vee 1)W_k\ . \tag{2.3}$$

Here $\{\psi_k(\cdot)\}$ are Borel functions from $\mathbb{R}^d$ to $\mathbb{R}^d$, $\{\eta_k\}$ is a sequence of $\mathbb{R}^d$-valued random variables, and $\{W_k\}$, $\{a_k\}$, $\{b_k\}$ are as in Section 2.1. Below we give sufficient

[+] The results are not changed if we replace $|X_k| \vee 1$ by $|X_k| \vee a$ or $|X_k| + a$ for $a \geq 0$.

conditions for the tightness of $\{X_k^x: k \geq 0, x \in K\}$ where K is a compact subset of $\mathbb{R}^d$. Note that algorithm (2.3) is somewhat more general than algorithm (2.1). The reason for considering this more general algorithm is that it is sometimes convenient to write an algorithm in the form (2.3) (with $\psi_k(x) \neq \nabla U(x)$ for some x, k) to verify tightness, and then to write the algorithm in the form (2.1) to verify convergence. We will give an example of this situation when we consider continuous state Metropolis-type annealing algorithms in Sections 3 and 4.

Let $\mathcal{G}_k = \sigma(X_0, W_0, ..., W_{k-1}, \eta_0, ..., \eta_{k-1})$. In the sequel we will consider the following conditions (α, β, γ_1, γ_2 are constants whose values will be specified later).

(B1) Let K be a compact subset of $\mathbb{R}^d$. Then

$$\sup_{k;x\in K} |\psi_k(x)| < \infty$$

$$\overline{\lim_{k,\,|x|\to\infty}} \frac{|\psi_k(x)|}{|x|} a_k^{\gamma_1} < \infty$$

$$\underline{\lim_{k,\,|x|\to\infty}} \frac{|\psi_k(x)|}{|x|} a_k^{\gamma_2} > 0$$

$$\underline{\lim_{k,\,|x|\to\infty}} \left\langle \frac{\psi_k(x)}{|\psi_k(x)|}, \frac{x}{|x|} \right\rangle > 0$$

(B2) There exists $L \geq 0$ such that

$$E\{|\eta_k|^2 \,|\mathcal{G}_k\} \leq L a_k^{\alpha}(|X_k|^2 \vee 1) \quad \text{w.p.1}$$
$$|E\{\eta_k \,|\mathcal{G}_k\}| \leq L a_k^{\beta}(|X_k| \vee 1) \quad \text{w.p.1}$$

W_k is independent of $\mathcal{G}_k$.

Theorem 2: Assume that (B1), (B2) hold with $\alpha > -1$, $\beta > 0$, and $0 \leq \gamma_2 \leq \gamma_1 < 1/2$. Let $\{X_k\}$ be given by (2.3) and K be a compact subset of $\mathbb{R}^d$. Then $\{X_k^x: k \geq 0, x \in K\}$ is a tight family of random variables.

Theorem 2 is proved similarly to [15, Theorem 3] where we considered the algorithm

$$X_{k+1} = X_k - a_k(\psi_k(X_k) + \eta_k) + b_k W_k \tag{2.4}$$

and we will not go through the details here. The main difference between the

conditions and proofs of Theorem 2 and [15, Theorem 3] is that in [15, Theorem 3] we allowed $\{\psi_k(x)\colon x \in \mathbb{R}^d\}$ to be a random vector field but we did not allow the bounds in (B2) to depend on $|x|$.

3. METROPOLIS-TYPE ANNEALING ALGORITHMS

In this Section we review the finite state Metropolis-type Markov chain annealing algorithm, generalize it to an arbitrary state space, and then specialize it to a class of algorithms for which the results in Section 2 can be applied to establish convergence.

The finite state Metropolis-type annealing algorithm may be described as follows [5]. Assume that the state space Σ is finite set. Let $U(\cdot)$ be a real valued function on Σ (the "energy" function) and $\{T_k\}$ be a sequence of strictly positive numbers (the "temperature" sequence). Let $q(i,j)$ be a stationary transition probability from i to j, for $i,j \in \Sigma$. The one-step transition probability at time k for the finite state Metropolis-type annealing chain $\{X_k\}$ is given by

$$\begin{aligned} P\{X_{k+1} = j \,|X_k = i\} &= q(i,j)s_k(i,j)\,, \qquad j \neq i\,, \\ P\{X_{k+1} = i \,|X_k = i\} &= 1 - \sum_{j \neq i} q(i,j)s_k(i,j) \end{aligned} \tag{3.1}$$

where

$$s_k(i,j) = \exp\left(-\frac{[U(j) - U(i)]_+}{T_k}\right). \tag{3.2}$$

This nonstationary Markov chain may be interpreted (and simulated) in the following manner. Given the current state $X_k = i$, generate a candidate state $\tilde{X}_k = j$ with probability $q(i,j)$. Set the next state $X_{k+1} = j$ if $s_k(i,j) > \theta_k$ where θ_k is an independent random variable uniformly distributed on the interval $[0,1]$; otherwise set $X_{k+1} = i$. Suppose that the stochastic matrix $Q = [q(i,j)]$ is symmetric and irreducible, and the temperature T_k is fixed at a constant $T > 0$. Then it can be shown that the resulting stationary Markov chain has a unique invariant Gibbs distribution with mass $\propto \exp(-U(i)/T)$, and furthermore converges to this Gibbs distribution as $k \to \infty$ [21]. There has been alot of work on the convergence and asymptotic behavior of the nonstationary annealing chain when $T_k \to 0$ [4]-[9].

We next generalize the finite state Metropolis-type annealing algorithm (3.1), (3.2) to a general state space. Assume that the state space Σ is a σ-finite measure space $(\Sigma, \wedge\ , \mu)$. Let $U(\cdot)$ be a real-valued measurable function on Σ and let $\{T_k\}$ be as above. Let $q(x,y)$ be a stationary transition probability density w.r.t. μ from x to y, for $x,y \in \Sigma$. The one-step transition probability at time k for the general state Metropolis-type annealing chain $\{X_k\}$ is given by

$$P\{X_{k+1} \in A \,|X_k = x\} = \int_A q(x,y)s_k(x,y)d\mu(y) + r_k(x)1_A(x) \tag{3.3}$$

where

$$r_k(x) = 1 - \int q(x,y)s_k(x,y)d\mu(y)\,, \tag{3.4}$$

and

$$s_k(x,y) = \exp\left(-\frac{[U(y) - U(x)]_+}{T_k}\right), \tag{3.5}$$

Note that if μ does not have an atom at x, then $r_k(x)$ is the self transition probability starting at state x at time k. Also note that (3.3)-(3.5) reduces to (3.1), (3.2) when the state space is finite and μ is counting measure. The general state chain may be interpreted (and simulated) similarly to the finite state chain: here, $q(x,y)$ is a conditional probability density for generating a candidate state $\tilde{X}_k = y$ given the current state $X_k = x$. Suppose that the stochastic transition function $Q(x,A) = \int_A q(x,y)d\mu(y)$ is μ-symmetric and irreducible, and the temperature T_k is fixed at a constant $T > 0$. Then similarly to the finite state case it can be shown that the resulting stationary Markov chain has a μ -a.e. unique invariant Gibbs distribution with density $\propto \exp(-U(x)/T)$, and furthermore if a certain condition due to Doeblin [21] is satisfied converges to this Gibbs distribution as $k \longrightarrow \infty$. There has been almost no work on the convergence and asymptotic behavior of the nonstationary annealing chain when $T_k \longrightarrow 0$, although when Σ is a compact metric space one would expect the behavior to be similar to when Σ is finite.

We next specialize the general state Metropolis-type annealing algorithm (3.3)-(3.5) to a d-dimensional Euclidean state space. Actually the Metropolis-type annealing chain we shall consider is not exactly a specialization of the general-state chain described above. Motivated by our desire to show convergence of the chain by writing it in the form of the modified stochastic gradient algorithm (2.1), we are led to choosing a

nonstationary Gaussian transition density

$$q_k(x,y) = \frac{1}{(2\pi b_k^2(|x|^2 \vee 1))^{d/2}} \exp\left(-\frac{1}{2}\frac{|y-x|^2}{b_k^2(|x|^2 \vee 1)}\right), \tag{3.6}$$

and a state dependent temperature sequence

$$T_k(x) = \frac{b_k^2(|x|^2 \vee 1)}{2a_k} \left(= \frac{\text{const.}(|x|^2 \vee 1)}{\log\log k}\right). \tag{3.7}$$

The choice of the transition density is clear, given we want to write the chain in the form of (2.1). The choice of the temperature sequence is based on the following considerations. Ignore for the moment the dependence on $|x|$ and examine the modified stochastic gradient algorithm (1.3) and the associated diffusion (1.1). If we view (1.3) as a sampled version of (1.1) with sampling intervals a_k and sampling times $t_k = \sum_{n=0}^{k-1} a_n$, then we have corresponding sampled temperatures $T(t_k) = c^2(t_k)/2$, and it is straightforward to check that if $C = B/A$ then

$$T_k = \frac{b_k^2}{2a_k} \sim \frac{c^2(t_k)}{2} = T(t_k) \text{ as } k \to \infty .$$

Finally, the fundamental reason that the $|x|$ dependence is needed in both (3.6), (3.7) is that in order to establish tightness of the annealing chain by writing the chain in either the form of (2.3) or (2.4) we need a condition like

$$|\psi_k(x)| \geq \text{const.}\,|x|\,, \quad |x| \text{ large}\,, \tag{3.8}$$

for suitable choice of $\psi_k(\cdot)$. In words, the annealing chain must generate a drift (towards the origin) at least proportional to the distance from the origin. To accomplish this we include the dependence on $|x|$ in (3.6), (3.7) and then write the chain in the form of (2.3) to establish tightness. This discussion leads us to the following continuous state Metropolis-type Markov chain annealing algorithm.

Metropolis-type Annealing Algorithm #1:

Let $\{X_k\}$ be a Markov chain with 1-step transition probability at time k given by[+]

$$P\{X_{k+1} \in A \,|X_k = x\} = \int_A s_k(x,y)dN(x, b_k^2(|x|^2 \vee 1)I)(y) + r_k(x)1_A(x) \qquad (3.9)$$

where

$$r_k(x) = 1 - \int s_k(x,y)dN(x, b_k^2(|x|^2 \vee 1)I)(y) \qquad (3.10)$$

and

$$s_k(x,y) = \exp\left(-\frac{2a_k}{b_k^2}\frac{[U(y) - U(x)]_+}{|x|^2 \vee 1}\right) \qquad (3.11)$$

Theorem 3: Assume (A1), (A2) hold and also

$$\sup_x |HU(x)| < \infty . \qquad (3.12)$$

Let $\{X_k\}$ be the Markov chain with transition probability given by (3.9)-(3.11). Then for $B/A > C_0$ and any bounded continuous function $f(\cdot)$ on $\mathbb{R}^d$

$$\lim_{k\to\infty} E_{0,x}\{f(X_k)\} = \pi(f) \qquad (3.13)$$

uniformly for x in a compact set.

The proof of Theorem 3 is in Section 4.1. Observe that the condition (3.12) can be rather restrictive. It implies along with (A1) that there exists constants M_1, M_2 such that

$$M_1|x| \le |\nabla U(x)| \le M_2|x|, \quad |x| \text{ large} .$$

It turns out that the lower bound on $|\nabla U(x)|$ is essential but the upper bound on $|\nabla U(x)|$ can be weakened by using a suitable modification of (3.11) as follows.

[+] The results are not changed if we replace $|x|^2 \vee 1$ by $|x|^2 \vee a$ or $|x|^2 + a$ for $a \ge 1$.

Metropolis-type Annealing Algorithm #2:

Let $\{X_k\}$ be a Markov chain with 1-step transition probability at time k given by[+]

$$P\{X_{k+1} \in A \,|X_k = x\} = \int_A s_k(x,y)dN(x, b_k^2(|x|^2 \vee 1)I)(y) + r_k(x)1_A(x) \quad (3.14)$$

where

$$r_k(x) = 1 - \int s_k(x,y)dN(x, b_k^2(|x|^2 \vee 1)I)(y) \quad (3.15)$$

and

$$s_k(x,y) = \exp\left(-\frac{2a_k}{b_k^2}\frac{[U(y) - U(x)]_+}{|x|^2 \vee 1}\right) \text{ if } U(x) \leq \frac{|x|^2 \vee 1}{a_k^\gamma}$$

$$= \exp\left(-\frac{2a_k}{b_k^2}\frac{[\,|y|^2 - |x|^2]_+}{|x|^2 \vee 1}\right) \text{ if } U(x) > \frac{|x|^2 \vee 1}{a_k^\gamma} \quad (3.16)$$

and $\gamma > 0$.

Theorem 4: Assume (A1), (A2) hold and also

$$\inf_{\delta>0} \overline{\lim_{|x|\to\infty}} \sup_{|y-x|<\delta|x|} |HU(y)| \cdot \frac{|x|^2}{U(x)} < \infty\,. \quad (3.17)$$

Let $\{X_k\}$ be the Markov chain with transition probability given by (3.14)-(3.16) with $0 < \gamma < 1/4$. Then for $B/A > C_0$ and any bounded continuous function $f(\cdot)$ on $\mathbb{R}^d$

$$\lim_{k\to\infty} E_{0,x}\{f(X_k)\} = \pi(f) \quad (3.18)$$

uniformly for x in a compact set.

The proof of Theorem 4 is in Section 4.2. Observe that the condition (3.17) (and also (A1)) will be satisfied if $U(x) \sim$ const. $|x|^p$ and $HU(x) = O(|x|^{p-2})$ as $|x| \to \infty$ for any $p \geq 2$. Note that if K is any fixed compact, $X_k \in K$, and k is very large, then (3.16) and (3.11) coincide. Note also that (3.16) like (3.11) only uses measurements of $U(\cdot)$ (and not $\nabla U(\cdot)$).

[+] The results are not changed if we replace $|x|^2 \vee 1$ by $|x|^2 \vee a$ or $|x|^2 + a$ for $a \geq 1$.

4. PROOFS OF THEOREMS 3 AND 4

In the sequel $c_1, c_2, \ldots$ will denote positive constants whose value may change from proof to proof. We will need the following lemma.

Lemma 1: Assume that $V(\cdot)$ is a C^2 function from $\mathbb{R}^d$ to $\mathbb{R}$. Let

$$s(x,y) = \exp(-\lambda[V(y) - V(x)]_+)$$

and

$$\hat{s}(x,y) = \exp(-\lambda[\langle \nabla V(x), y-x \rangle \ ]_+)$$

where $\lambda > 0$. Then

$$|s(x,y) - \hat{s}(x,y)| \leq \lambda \sup_{\epsilon \in (0,1)} |HV(x + \epsilon(y-x))| \, |y-x|^2$$

for all $x,y \in \mathbb{R}^d$.

Proof: Let

$$f(x,y) = V(y) - V(x) - \langle \nabla V(x), y-x \rangle \quad .$$

Then by the 2nd order Taylor Theorem

$$|f(x,y)| \leq \sup_{\epsilon \in (0,1)} |HV(x + \epsilon(y-x))| \, |y-x|^2 \tag{4.1}$$

By separately considering the four cases corresponding to the possible signs of $V(y) - V(x)$ and $\langle \nabla V(x), y-x \rangle$, it can be shown that

$$|s(x,y) - \hat{s}(x,y)| \leq 1 - \exp(-\lambda |f(x,y)|) \leq \lambda |f(x,y)| \tag{4.2}$$

Combining (4.1) and (4.2) completes the proof.

□

4.1. Proof of Theorem 3

We write

$$X_{k+1} = X_k - a_k(\nabla U(X_k) + \xi_k) + b_k(|X_k| \vee 1)W_k \tag{4.3}$$

(this defines ξ_k) and apply Theorem 1 to show that if $\{X_k^x: k \geq 0, x \in K\}$ is tight for K compact then (3.13) is true. We further let $\psi_k(x) = \nabla U(x)$ and $\eta_k = \xi_k$ and apply Theorem 2 to show that $\{X_k^x: k \geq 0, x \in K\}$ is infact tight for K compact and (3.13) is

infact true.

We first show that we can find a version of $\{X_k\}$ in the form

$$X_{k+1} = X_k + b_k(|X_k| \vee 1)\varsigma_k W_k \tag{4.4}$$

To do this we inductively define the sequence $\{W_k, \varsigma_k\}$ of random variables as follows. Assume $X_0, W_0, \cdots, W_{k-1}, \varsigma_0, \cdots, \varsigma_{k-1}$ have been defined. Let $\mathcal{A}_k = \sigma(X_0, W_0, W_{k-1}, \varsigma_0, \cdots, \varsigma_{k-1})$. Let W_k be a standard d-dimensional Gaussian random variable independent of $\mathcal{A}_k$, and let ς_k be a $\{0,1\}$-valued random variable with

$$P\{\varsigma_k = 1 \,|\mathcal{A}_k, W_k\} = s_k(X_k, X_k + b_k(|X_k| \vee 1)W_k)\,. \tag{4.5}$$

Note that $P\{\varsigma_k = i \,|\mathcal{A}_k, W_k) = P\{\varsigma_k = i \,|X_k, W_k\}$. Using (4.5) it is easy to check that (4.4) is a Markov chain which has transition probability given by (3.9)-(3.11). Hence (4.4) is indeed a version of $\{X_k\}$ and we always deal with this version in the sequel.

Now comparing (4.3) and (4.4) we have

$$\xi_k = -\nabla U(X_k) + \frac{b_k}{a_k}(|X_k| \vee 1)(1 - \varsigma_k)W_k \tag{4.6}$$

and in particular ξ_k is a function of X_k, W_k and ς_k. Note that since $\mathcal{F}_k \subset \mathcal{A}_k$, W_k is independent of $\mathcal{A}_k$, and $P\{\varsigma_k = i \,|\mathcal{A}_k, W_k\} = P\{\varsigma_k = i \,|X_k, W_k\}$, it follows that W_k is independent of $\mathcal{F}_k$ and $P\{\varsigma_k = i \,|\mathcal{F}_k, W_k\} = P\{\varsigma_k = i \,|X_k, W_k\}$. Hence $P\{\xi_k \in A \,|\mathcal{F}_k\} = P\{\xi_k \in A \,|X_k\}$. We will use these facts below.

The following lemma gives the crucial estimates for $E\{|\xi_k|^2 \,|\mathcal{F}_k\}$ and $|E\{\xi_k \,|\mathcal{F}_k\}|$.

Lemma 2: There exists $L \geq 0$ such that

a) $|E\{\xi_k \,|\mathcal{F}_k\}| \leq L\frac{a_k}{b_k}(|X_k| \vee 1)$ w.p.1

b) $E\{|\xi_k|^2 \,|\mathcal{F}_k\} \leq L\frac{b_k}{a_k}(|X_k|^2 \vee 1)$ w.p.1

Assume that Lemma 2 is true. Then (A3) is satisfied with $\alpha = -\frac{1}{2} > -1$ and $0 < \beta < \frac{1}{2}$, and (B1), (B2) are satisfied for the same choice of α and β and for $\gamma_1 = \gamma_2 = 0$. Hence Theorems 1 and 2 apply and Theorem 3 follows. It remains to

prove Lemma 2. We will use the following claim.

Claim: Let $u \in \mathbb{R}^d$ with $|u| = 1$. Then

a) $$\int_{0 \le \langle u,w \rangle \le \delta} dN(0,I)(w) = O(\delta)$$

b) $$\int_{0 \le \langle u,w \rangle \le \delta} w dN(0,I)(w) = O(\delta^2)$$

c) $$\int_{0 \le \langle u,w \rangle \le \delta} w \otimes w dN(0,I)(w) = O(\delta)$$

Proof: Let $u_1 = u$ and extend u_1 to an orthonormal basis $\{u_1,...,u_d\}$ for $\mathbb{R}^d$. Then by changing variables (rotation) and using the Mean Value Theorem we get

a) $$\int_{0 \le \langle u,w \rangle \le \delta} dN(0,I)(w) = \int_0^\delta \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{v^2}{2}\right) dv = O(\delta)$$

b) $$\int_{0 \le \langle u,w \rangle \le \delta} w dN(0,I)(w) = u_1 \int_0^\delta v \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{v^2}{2}\right) dv = O(\delta^2)$$

c) $$\int_{0 \le \langle u,w \rangle \le \delta} w \otimes w dN(0,I)(w) = u_1 \otimes u_1 \int_0^\delta v^2 \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{v^2}{2}\right) dv$$

$$+ \sum_{i=2}^{d} u_i \otimes u_i \int_0^\delta \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{v^2}{2}\right) dv$$

$$= O(\delta)$$

□

Proof of Lemma 2a):

Using (4.6) and the fact that $P\{\xi_k \in A \,|\mathcal{F}_k\} = P\{\xi_k \in A \,|X_k\}$ and W_k is independent of X_k we have (w.p.1)

$$
\begin{aligned}
E\{\xi_k \,|\mathcal{F}_k\} &= E\{\xi_k \,|X_k\} \\
&= -\nabla U(X_k) + \frac{b_k}{a_k}(\,|X_k\,| \vee 1)E\{(1-\varsigma_k)W_k \,|X_k\} \\
&= -\nabla U(X_k) - \frac{b_k}{a_k}(\,|X_k\,| \vee 1)E\{W_k E\{\varsigma_k \,|X_k, W_k\} \,|X_k\} \\
&= -\nabla U(X_k) - \frac{b_k}{a_k}(\,|X_k\,| \vee 1)E\{W_k P\{\varsigma_k = 1 \,|X_k, W_k\} \,|X_k\} \\
&= -\nabla U(X_k) - \frac{b_k}{a_k}(\,|X_k\,| \vee 1)\underset{W_k}{E}\{W_k P\{\varsigma_k = 1 \,|X_k, W_k\}\}\,.
\end{aligned}
$$

Henceforth we condition on $X_k = x$ where $|x| \geq 1$; the case where $|x| \leq 1$ is similar. Hence using (4.5)

$$
E\{\xi_k \,|X_k = x\} = -\nabla U(x) - \frac{b_k}{a_k}\,|x| \int w s_k(x, x + b_k\,|x|\,w) dN(0,I)(w)\,.
$$

Let

$$
\hat{s}_k(x,y) = \exp\left[-\frac{2a_k}{b_k^2}\frac{[\langle \nabla U(x), y-x\rangle\ ]_+}{|x|^2}\right] \tag{4.7}
$$

and $\tilde{s}_k(x,y) = s_k(x,y) - \hat{s}_k(x,y)$. Then by (3.12) and Lemma 1

$$
|\tilde{s}_k(x,y)| \leq c_1 \frac{a_k}{b_k^2}\frac{|y-x|^2}{|x|^2}\,. \tag{4.8}
$$

Hence

$$E\{\xi_k \,|X_k=x\} = -\nabla U(x) - \frac{b_k}{a_k}\,|x\,|\int w\hat{s}_k(x,x+b_k\,|x\,|w)dN(0,I)(w)$$

$$- \frac{b_k}{a_k}\,|x\,|\int w\tilde{s}_k(x,x+b_k\,|x\,|w)dN(0,I)(w)$$

$$= -\nabla U(x) - \frac{b_k}{a_k}\,|x\,|\int w\hat{s}_k(x,x+b_k\,|x\,|w)dN(0,I)(w)$$

$$+ O(b_k\,|x\,|) \tag{4.9}$$

$$= -\nabla U(x) - \frac{b_k}{a_k}\,|x\,| \int\limits_{\langle\, \nabla U(x),w\rangle\, \le 0} w dN(0,I)(w)$$

$$- \frac{b_k}{a_k}\,|x\,| \int\limits_{\langle\, \nabla U(x),w\rangle\, > 0} w \exp\left(-\frac{2a_k}{b_k}\frac{\langle \nabla U(x),w\rangle}{|x\,|}\right) dN(0,I)(w)$$

$$+ O(b_k\,|x\,|)\,. \tag{4.10}$$

Clearly

$$E\{\xi_k\,|X_k = x\} = O(b_k\,|x\,|) \tag{4.11}$$

for x such that $\nabla U(x) = 0$. Henceforth we assume that $\nabla U(x) \neq 0$. Let $\nabla\hat{U}(x) = \nabla U(x)/|\nabla U(x)\,|$. Completing the square in the second integral in (4.10) we get

$$E\{\xi_k\,|X_k = x\} = -\nabla U(x) - \frac{b_k}{a_k}\,|x\,| \int\limits_{\langle\, \nabla\hat{U}(x),w\rangle\, \le 0} w dN(0,I)(w)$$

$$- \frac{b_k}{a_k}\,|x\,| \int\limits_{\langle\, \nabla\hat{U}(x),w\rangle\, \ge\, 0} w \exp\left(2\left(\frac{a_k}{b_k}\right)^2\left(\frac{|\nabla U(x)\,|}{|x\,|}\right)^2\right) dN\left(-\frac{2a_k}{b_k}\frac{\nabla U(x)}{|x\,|}\,,\, I\right)(w)$$

$$+ O(b_k\,|x\,|) \tag{4.12}$$

Now by (3.12) $|\nabla U(x)\,| = O(|x\,|)$ and so

$$\exp\left(2\left(\frac{a_k}{b_k}\right)^2\left(\frac{|\nabla U(x)\,|}{|x\,|}\right)^2\right) = 1 + O\left(\left(\frac{a_k}{b_k}\right)^2\right). \tag{4.13}$$

Substituting (4.13) into (4.12), using $a_k/b_k = O(1)$ and $|\nabla U(x)\,| = O(|x\,|)$, and changing variables from $w + 2(a_k/b_k)(\nabla U(x)/|x\,|)$ to w gives

$$E\{\xi_k \,|X_k = x\} = -\nabla U(x) - \frac{b_k}{a_k}|x| \int_{\langle \nabla\hat{U}(x),w\rangle \le 0} w dN(0,I)(w)$$

$$- \frac{b_k}{a_k}|x| \int_{\langle \nabla\hat{U}(x),w\rangle \ge O(\frac{a_k}{b_k})} w dN(0,I)(w) + 2\nabla U(x) \int_{\langle \nabla\hat{U}(x),w\rangle \ge O(\frac{a_k}{b_k})} dN(0,I)(w)$$

$$+ O\left(\frac{a_k}{b_k}|x|\right) + O(b_k|x|)$$

$$= \frac{b_k}{a_k}|x| \int_{0 \le \langle \nabla\hat{U}(x),w\rangle \le O(\frac{a_k}{b_k})} w dN(0,I)(w)$$

$$- 2\nabla U(x) \int_{0 \le \langle \nabla\hat{U}(x),w)\rangle \le O(\frac{a_k}{b_k})} dN(0,I)(w)$$

$$+ O\left(\frac{a_k}{b_k}|x|\right) \tag{4.14}$$

Hence by the Claim parts a), b) and again using $|\nabla U(x)| = O(|x|)$ we have

$$E\{\xi_k \,|X_k = x\} = O\left(\frac{a_k}{b_k}|x|\right) \tag{4.15}$$

Combining (4.11) and (4.15) completes the proof of Lemma 2a).

□

Proof of Lemma 2b):

Using (4.6) and the fact that $P\{\xi_k \in A \,|\mathcal{F}_k\} = P\{\xi_k \in A \,|X_k\}$, W_k is independent of X_k, and $|\nabla U(x)| = O(|x|)$ we have (w.p.1)

$$E\{\xi_k \otimes \xi_k \,|\mathcal{F}_k\} = E\{\xi_k \otimes \xi_k \,|X_k\}$$

$$= \left(\frac{b_k}{a_k}\right)^2 (\,|X_k|^2 \vee 1) E\{((1-\zeta_k)W_k) \otimes ((1-\zeta_k)W_k)\,|X_k\} + e_k(X_k)$$

$$= \left(\frac{b_k}{a_k}\right)^2 (\,|X_k|^2 \vee 1) I - \left(\frac{b_k}{a_k}\right)^2 (\,|X_k|^2 \vee 1) E\{W_k \otimes W_k E\{\zeta_k \,|X_k, W_k\}\,|X_k\} + e_k(X_k)$$

$$= \left(\frac{b_k}{a_k}\right)^2 (\,|X_k|^2 \vee 1) I - \left(\frac{b_k}{a_k}\right)^2 (\,|X_k|^2 \vee 1) E\{W_k \otimes W_k P\{\zeta_k = 1 \,|X_k, W_k\}\,|X_k\} + e_k(X_k)$$

$$= \left(\frac{b_k}{a_k}\right)^2 (\,|X_k|^2 \vee 1) I - \left(\frac{b_k}{a_k}\right)^2 (\,|X_k|^2 \vee 1) \underset{W_k}{E}\{W_k \otimes W_k P\{\zeta_k = 1 \,|X_k, W_k\}\} + e_k(X_k)$$

where

$$e_k(X_k) = O\left(\frac{b_k}{a_k}(\,|X_k|^2 \vee 1)\right).$$

Henceforth we condition on $X_k = x$ where $|x| \geq 1$; the case where $|x| \leq 1$ is similar. Hence using (4.5)

$$E\{\xi_k \otimes \xi_k \,|X_k = x\} = \left(\frac{b_k}{a_k}\right)^2 |x|^2 I - \left(\frac{b_k}{a_k}\right)^2 |x|^2 \int w \otimes w s_k(x, x + b_k\,|x|w) dN(0,I)(w)$$

$$+ O\left(\frac{b_k}{a_k}\,|x|^2\right).$$

Let $\hat{s}_k(x,y)$ be given by (4.7) and $\tilde{s}_k(x,y) = s_k(x,y) - \hat{s}_k(x,y)$. Then using (4.8)

$$E\{\xi_k \otimes \xi_k \,|X_k = x\} = \left(\frac{b_k}{a_k}\right)^2 |x|^2 I - \left(\frac{b_k}{a_k}\right)^2 |x|^2 \int w \otimes w\hat{s}_k(x, x + b_k |x| w) dN(0,I)(w)$$

$$- \left(\frac{b_k}{a_k}\right)^2 |x|^2 \int w \otimes w\tilde{s}_k(x, x + b_k |x| w) dN(0,I)(w)$$

$$+ O\left(\frac{b_k}{a_k} |x|^2\right)$$

$$= \left(\frac{b_k}{a_k}\right)^2 |x|^2 I - \left(\frac{b_k}{a_k}\right)^2 |x|^2 \int w \otimes w\hat{s}_k(x, x + b_k |x| w) dN(0,I)(w)$$

$$+ O\left(\frac{b_k^2}{a_k} |x|^2\right) + O\left(\frac{b_k}{a_k} |x|^2\right) \tag{4.16}$$

$$= \left(\frac{b_k}{a_k}\right)^2 |x|^2 I - \left(\frac{b_k}{a_k}\right)^2 |x|^2 \int_{\langle \nabla U(x), w \rangle \leq 0} w \otimes w dN(0,I)(w)$$

$$- \left(\frac{b_k}{a_k}\right)^2 |x|^2 \int_{\langle \nabla U(x), w \rangle > 0} w \otimes w \exp\left(-\frac{2a_k}{b_k} \frac{\langle \nabla U(x), w \rangle}{|x|}\right) dN(0,I)(w)$$

$$+ O\left(\frac{b_k}{a_k} |x|^2\right). \tag{4.17}$$

Clearly

$$E\{\xi_k \otimes \xi_k \,|X_k = x\} = O\left(\frac{b_k}{a_k} |x|^2\right) \tag{4.18}$$

for x such that $\nabla U(x) = 0$. Henceforth we assume that $\nabla U(x) \neq 0$. Let $\nabla \hat{U}(x) = \nabla U(x) / |\nabla U(x)|$. Completing the square in the second integral in (4.17), and proceeding similiarly to the derivation of (4.14) in the proof of Lemma 2a) we have

$$E\{\xi_k \otimes \xi_k \,|X_k = x\} = \left(\frac{b_k}{a_k}\right)^2 |x|^2 I - \left(\frac{b_k}{a_k}\right)^2 |x|^2 \int_{\langle \hat{U}(x), w\rangle \le 0} w \otimes w dN(0,I)(w)$$

$$- \left(\frac{b_k}{a_k}\right)^2 |x|^2 \int_{\langle \nabla\hat{U}(x), w\rangle \ge 0} w \otimes w \exp\left(2\left(\frac{a_k}{b_k}\right)^2 \left(\frac{|\nabla U(x)|}{|x|}\right)^2\right) dN\left(-\frac{2a_k}{b_k}\frac{\nabla U(x)}{|x|}, I\right)(w)$$

$$+ O\left(\frac{b_k}{a_k}|x|^2\right)$$

$$= \left(\frac{b_k}{a_k}\right)^2 |x|^2 I - \left(\frac{b_k}{a_k}\right)^2 |x|^2 \int_{\langle \nabla\hat{U}(x), w\rangle \le 0} w \otimes w dN(0,I)(w)$$

$$- \left(\frac{b_k}{a_k}\right)^2 |x|^2 \int_{\langle \nabla\hat{U}(x), w\rangle \ge O(\frac{a_k}{b_k})} w \otimes w dN(0,I)(w)$$

$$+ O(|x|^2) + O\left(\frac{b_k}{a_k}|x|^2\right)$$

$$= \left(\frac{b_k}{a_k}\right)^2 |x|^2 \int_{0 \le \langle \nabla\hat{U}(x), w\rangle \le O(\frac{a_k}{b_k})} w \otimes w dN(0,I)(w)$$

$$+ O\left(\frac{b_k}{a_k}|x|^2\right)$$

Hence by the Claim part c)

$$E\{\xi_k \otimes \xi_k \,|X_k = x\} = O\left(\frac{b_k}{a_k}|x|^2\right) \tag{4.19}$$

Combining (4.18) and (4.19) and using the fact that $|\xi_k \otimes \xi_k| \le |\xi_k|^2$ completes the proof of Lemma 2b).

□

Remark: In Figure 1 we demonstrate the type of approximations used in the proof of Theorem 3. In Figure 1(i) we show the transition density $p_k(x,y)$ for the Markov chain with transition probability given by (3.9)-(3.11); in Figure 1(ii) we show the transition density $p'_k(x,y)$ for the same Markov chain but using $\hat{s}_k(x,y)$ (eqn. (4.7)) in place of $s_k(x,y)$ (eqn. (3.11)); and in Figure 1(iii) we show the transition density $p''_k(x,y)$ for the Markov chain of (2.1) with $\xi_k = 0$. Note that the densities in Figures 1(i) and (ii) contain impulsive components associated with the positive probability of no transition. All three densities are "close" for sufficiently large k.

4.2. Proof of Theorem 4

We write

$$X_{k+1} = X_k - a_k(\nabla U(X_k) + \xi_k) + b_k(|X_k| \vee 1)W_k$$

(this defines ξ_k) and apply Theorem 1 to show that if $\{X_k^x: k \geq 0, x \in K\}$ is tight for K compact then (3.18) is true. We further let

$$\psi_k(x) = \nabla U(x) \quad \text{if } U(x) \leq \frac{|x|^2 \vee 1}{a_k^\gamma}$$

$$= 2x \quad \text{if } U(x) > \frac{|x|^2 \vee 1}{a_k^\gamma}$$

and write

$$X_{k+1} = X_k - a_k(\psi_k(X_k) + \eta_k) + b_k(|X_k| \vee 1)W_k$$

(this defines η_k) and apply Theorem 2 to show that $\{X_k^x: k \geq 0, x \in K\}$ is infact tight for K compact and (3.18) is infact true.

The following lemmas give the crucial estimates for $E\{|\xi_k|^2 | \mathcal{F}_k\}$, $|E\{\xi_k | \mathcal{F}_k\}|$, $E\{|\eta_k|^2 | \mathcal{G}_k\}$ and $|E\{\eta_k | \mathcal{G}_k\}|$ (compare with Lemma 2).

Lemma 3: Let K be a compact subset of $\mathbb{R}^d$. Then there exists $L \geq 0$ such that

a) $|E\{\xi_k | \mathcal{F}_k\}| \leq L\frac{a_k}{b_k} \; \forall \, X_k \in K$, w.p.1

b) $E\{|\xi_k|^2 \,|\mathcal{F}_k\} \le L\frac{b_k}{a_k} \;\forall\, X_k \in K$, w.p.1

Lemma 4: There exists $L \ge 0$ such that

a) $|E\{\eta_k \,|\mathcal{G}_k\}| \le L\frac{a_k^{1-2\gamma}}{b_k}(|X_k| \vee 1)$ w.p.1

b) $E\{|\eta_k|^2 \,|\mathcal{G}_k\} \le L\frac{b_k}{a_k^{1+\gamma}}(|X_k|^2 \vee 1)$ w.p.1

Assume that Lemmas 3 and 4 are true. Then (A3) is satisfied with $\alpha = -\frac{1}{2} - \gamma > -1$ and $0 < \beta < \frac{1}{2} - 2\gamma$, and (B1), (B2) are satisfied with $\alpha = -\frac{1}{2}$, $0 < \beta < \frac{1}{2}$, $\gamma_1 = \gamma$ and $\gamma_2 = 0$ (recall that we assume $0 < \gamma < \frac{1}{4}$). Hence Theorems 1 and 2 apply, and Theorem 4 follows. It remains to prove Lemmas 3 and 4.

Proof of Lemma 3:

In the sequel we condition on $X_k = x$ where $x \in K$ and $|x| \ge 1$; the case where $|x| \le 1$ is similar. Let

$$\hat{s}_k(x,y) = \exp\left(-\frac{2a_k}{b_k^2}\frac{[\langle \nabla U(x), y-x\rangle\ ]_+}{|x|^2}\right) \text{ if } U(x) \le \frac{|x|^2}{a_k^\gamma}$$

$$= \exp\left(-\frac{2a_k}{b_k^2}\frac{[\langle 2x, y-x\rangle\ ]_+}{|x|^2}\right) \quad \text{if } U(x) > \frac{|x|^2}{a_k^\gamma} \tag{4.20}$$

and $\tilde{s}_k(x,y) = s_k(x,y) - \hat{s}_k(x,y)$. Using the fact that $HU(\cdot)$ is bounded on a compact we get for any fixed $\delta > 0$

$$\sup_{\epsilon\in(0,1)} |HU(x + \epsilon(y-x))| \le \sup_{|z-x|<\delta|x|} |HU(z)| \le c_1 ,$$

for all $|y-x| < \delta|x|$, and in particular the inequality holds when $U(z) = |z|^2$. Hence by considering the two cases where $U(x)$ is $\le$ or $>$ $|x|^2/a_k^\gamma$ and using Lemma 1 we get

$$|\tilde{s}_k(x,y)| \leq c_2 \frac{a_k}{b_k^2} \frac{|y-x|^2}{|x|^2} \quad , \quad |y-x| < \delta|x| \tag{4.21}$$

Note that (4.21) unlike (4.8) only holds for $|y-x| < \delta|x|$. Of course

$$|\tilde{s}_k(x,y)| \leq 1 \tag{4.22}$$

Using (4.21), (4.22) and a standard estimate for the tail probability of a Gaussian random variable we get for $i \geq 0$

$$\begin{aligned}
&\int |w|^i \, |\tilde{s}_k(x, x + b_k|x|w)| dN(0,I)(w) \\
&\leq \int_{|w| \leq \delta/b_k} |w|^i \, |\tilde{s}_k(x, x + b_k|x|w)| dN(0,I)(w) \\
&\quad + \int_{|w| > \delta/b_k} |w|^i \, |\tilde{s}_k(x, x + b_k|x|w)| dN(0,I)(w) \\
&\leq c_3 a_k + c_3 \exp\left(-\frac{c_4}{b_k^2}\right) \\
&= O(a_k)
\end{aligned} \tag{4.23}$$

Using (4.23) we get similarly to the derivation of (4.9) and (4.16)

$$\begin{aligned}
E\{\xi_k | X_k = x\} = -\nabla U(x) - \frac{b_k}{a_k}|x| \int w \hat{s}_k(x, x + b_k|x|w) dN(0,I)(w) \\
+ O(b_k)
\end{aligned} \tag{4.24}$$

and

$$\begin{aligned}
E\{\xi_k \otimes \xi_k | X_k = x\} = \left(\frac{b_k}{a_k}\right)^2 |x|^2 I - \left(\frac{b_k}{a_k}\right)^2 |x|^2 \int w \otimes w \hat{s}_k(x, x + b_k|x|w) dN(0,I)(w) \\
+ O\left(\frac{b_k}{a_k}\right)
\end{aligned} \tag{4.25}$$

Now recall that $x \in K$ and $\gamma > 0$. Hence for k large enough (and it is enough to consider large k), $U(x) \leq |x|^2/a_k^\gamma$ so that $\hat{s}_k(x,y)$ which was defined by (4.20) is the same as (4.7), and consequently (4.24) and (4.25) are the same equations as (4.9) and (4.16),

respectively (except for the error terms). Lemma 3 now follows by the same procedure as in the proof of Lemma 2 except now $\nabla U(x) = O(1)$ instead of $\nabla U(x) = O(|x|)$.

□

Proof of Lemma 4:

In the sequel we condition on $X_k = x$ where $|x| \geq 1$; the case where $|x| \leq 1$ is similar. Let $\hat{s}_k(x,y)$ be given by (4.20) and $\tilde{s}_k(x,y) = s_k(x,y) - \hat{s}_k(x,y)$. Using (3.17) we get for some $\delta > 0$

$$\sup_{\epsilon \in (0,1)} |HU(x + \epsilon(y-x))| \leq \sup_{|z-x| < \delta|x|} |HU(z)| \leq c_1 \left(\frac{U(x)}{|x|^2} + 1 \right),$$

for all $|y-x| < \delta|x|$, and in particular the inequality holds when $U(z) = |z|^2$. Hence by considering the two cases where $U(x)$ is $\leq$ or $>$ $|x|^2/a_k^\gamma$ and using Lemma 1 we get

$$|\tilde{s}_k(x,y)| \leq c_2 \frac{a_k^{1-\gamma}}{b_k^2} \frac{|y-x|^2}{|x|^2} \quad , \quad |y-x| < \delta|x| \, . \tag{4.26}$$

Using (4.26) we get similarly to the derivation of (4.23)

$$\int |w|^i \, |\tilde{s}_k(x, x + b_k |x| w)| \, |dN(0,I)(w) = O(a_k^{1-\gamma}) \tag{4.27}$$

Using (4.27) we get similiarly to the derivation of (4.9) and (4.16)

$$E\{\eta_k | X_k = x\} = -\psi_k(x) - \frac{b_k}{a_k} |x| \int w \hat{s}_k(x, x + b_k |x| w) dN(0,I)(w)$$

$$+ O\left(\frac{b_k}{a_k^\gamma} |x| \right) \tag{4.28}$$

and

$$E\{\eta_k \otimes \eta_k | X_k = x\} = \left(\frac{b_k}{a_k} \right)^2 |x|^2 I - \left(\frac{b_k}{a_k} \right)^2 |x|^2 \int w \otimes w \hat{s}_k(x, x + b_k |x| w) dN(0,I)(w)$$

$$+ O\left(\frac{b_k}{a_k^{1+\gamma}} |x|^2 \right) \tag{4.29}$$

Now (4.28) and (4.29) are the same equations as (4.9) and (4.16), respectively, with $\nabla U(x)$ replaced by $\psi_k(x)$ and ξ_k replaced by η_k (except for the error terms). Lemma 4

now follows by the same procedure as in the proof of Lemma 2 except now $\psi_k(x) = O(|x|/a_k^\gamma)$ instead of $\nabla U(x) = O(|x|)$.

□

5. REFERENCES

[1] Kirkpatrick, S., Gelatt C. D., and Vecchi, M., *Optimization by Simulated Annealing,* Science, Vol. 220, pp. 621-680, 1983.

[2] Cerny, V., *A Thermodynamical Approach to the Travelling Salesman Problem,* Journal of Optimization Theory and Applications, Vol. 45, pp. 41-51, 1985.

[3] Binder, K., *Monte Carlo Methods in Statistical Physics,* Springer Verlag, Berlin, 1978.

[4] Geman, S., and Geman, D., *Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images,* IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-6, pp. 721-741, 1984.

[5] Gidas, B., *Nonstationary Markov Chains and Convergence of the Annealing Algorithm,* Journal of Statistical Physics, Vol. 39, pp. 73-131, 1985.

[6] Hajek, B., *Cooling Schedules for Optimal Annealing,* Mathematics of Operations Research, Vol. 13, pp. 311-329, 1988.

[7] Mitra, D., Romeo, F., and Sangiovanni-Vincentelli, A., *Convergence and Finite-Time Behavior of Simulated Annealing,* Advances in Applied Probability, Vol. 18, pp. 747-771, 1986.

[8] Tsitsiklis, J., *Markov Chains with Rare Transitions and Simulated Annealing,* Mathematics of Operations Research, Vol. 14, pp. 70-90, 1989.

[9] Tsitsiklis, J., *A Survey of Large Time Asymptotics of Simulated Annealing Algorithms,* Massachusetts Institute of Technology, Laboratory for Information and Decision Systems, Report No. LIDS-P-1623, 1986.

[10] Geman, S. and Hwang, C. R., *Diffusions for Global Optimization,* SIAM Journal Control and Optimization, 24, pp. 1031-1043, 1986.

[11] Grenender, U., *Tutorial in Pattern Theory,* Div. Applied Mathematics, Brown Univ., Providence, RI, 1984.

[12] Chiang, T. S., Hwang, C. R. and Sheu, S. J., *Diffusion for Global Optimization in* $\mathbb{R}^n$, SIAM Journal Control and Optimization, 25, pp. 737-752, 1987.

[13] Gidas, B., *Global Optimization via the Langevin Equation,* Proc. IEEE Conference on Decision and Control, 1985.

[14] Kushner, H. J., *Asymptotic Global Behavior for Stochastic Approximation and Diffusions with Slowly Decreasing Noise Effects: Global Minimization via Monte Carlo,* SIAM Journal Applied Mathematics, 47, pp. 169-185, 1987.

[15] Gelfand, S. B. and Mitter, S. K., *Recursive Stochastic Algorithms for Global Optimization in $\mathbb{R}^d$,* submitted to SIAM Journal Control and Optimization, 1990.

[16] Kushner, H. J. and Clark, D., *Stochastic Approximation Methods for Constrained and Unconstrained Systems,* Springer, Berlin, 1978.

[17] Ljung, L., *Analysis of Recursive Stochastic Algorithms,* IEEE Transactions on Automatic Control, Vol. AC-22, pp. 551-575, 1977.

[18] Brook, D. G. and Verdini, W. A., *Computational Experience with Generalized Simulated Annealing Over Continuous Variables,* American Journal of Mathematical and Management Sciences, Vol. 8, Nos. 3 and 4, pp. 425-449, 1988.

[19] Jeng, F. C. and Woods, J. W., *Simulated Annealing in Compound Gaussian Random Fields,* IEEE Transactions on Information Theory, Vol. IT-36, pp. 94-107, 1990.

[20] Hwang, C. R., *Laplaces Method Revisited: Weak Convergence of Probability Measures,* Annals of Probability, 8, pp. 1177-1182, 1980.

[21] Doob, J. L., *Stochastic Processes,* Wiley, New York, 1953.

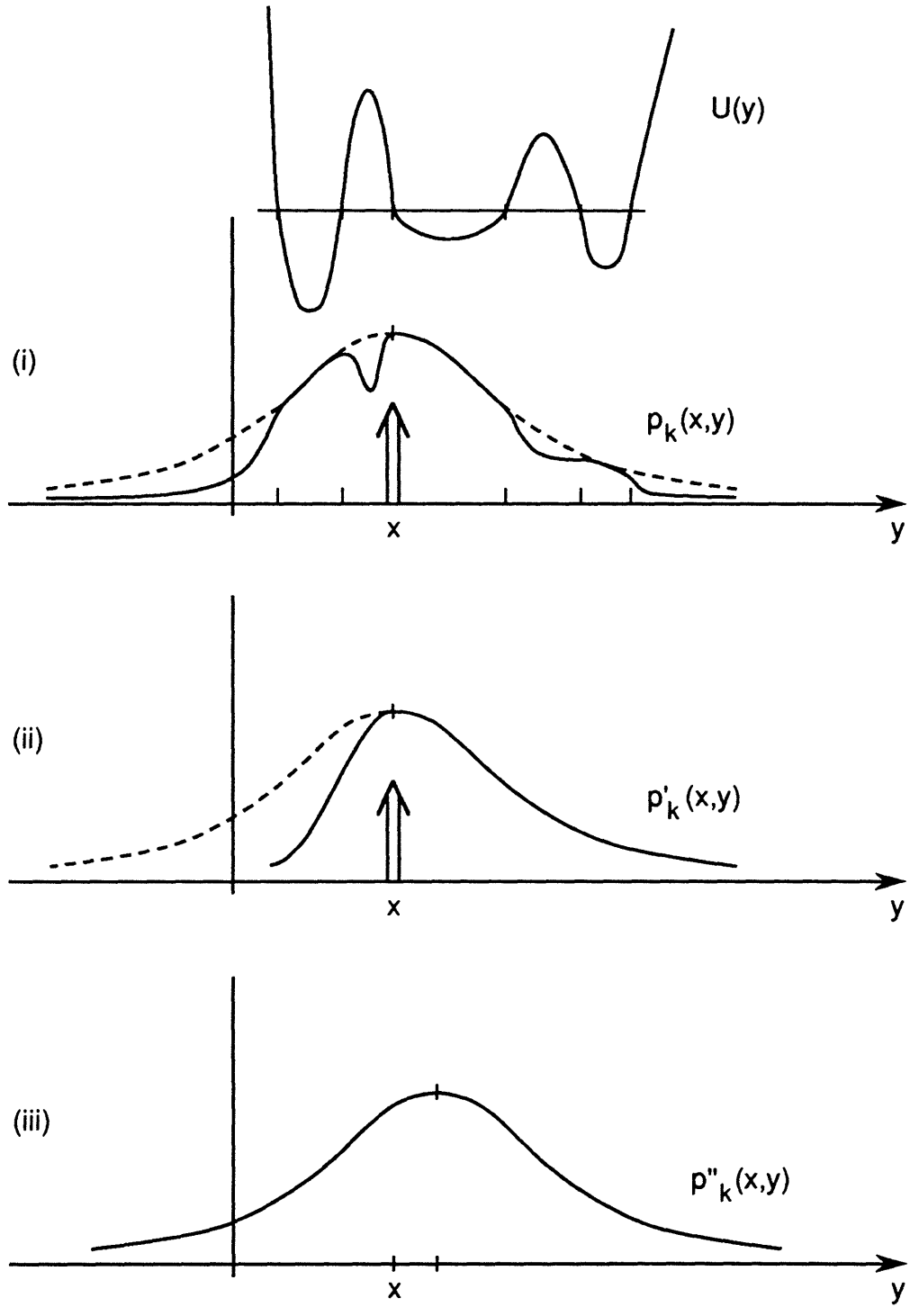

Figure 1. Three transition probability densities